

Attorney Docket No. 83996
Customer No. 23523

DETECTION OF RANDOMNESS IN SPARSE DATA SET OF
THREE DIMENSIONAL TIME SERIES DISTRIBUTIONS

TO WHOM IT MAY CONCERN:

BE IT KNOWN THAT FRANCIS J. O'BRIEN, JR, employee of the United States Government, citizen of the United States of America, resident of Newport, County of Newport, State of Rhode Island, has invented certain new and useful improvements entitled as set forth above of which the following is a specification:

MICHAEL F. OGLO, ESQ.
Reg. No. 20464
Naval Undersea Warfare Center
Division Newport
Newport, RI 02841-1708
TEL: 401-832-4736
FAX: 401-832-1231

1 Attorney Docket No. 83996

2

3 DETECTION OF RANDOMNESS IN SPARSE DATA SET OF

4 THREE DIMENSIONAL TIME SERIES DISTRIBUTIONS

5

6 STATEMENT OF GOVERNMENT INTEREST

7 The invention described herein may be manufactured and used
8 by or for the Government of the United States of America for
9 Governmental purposes without the payment of any royalties
10 thereon or therefore.

11

12 BACKGROUND OF THE INVENTION

13 (1) Field of the Invention

14 The invention generally relates to signal processing/data
15 processing systems for processing time series distributions
16 containing a small number of data points (e.g., less than about
17 ten (10) to twenty-five (25) data points). More particularly,
18 the invention relates to a dual method for classifying the white
19 noise degree (randomness) of a selected signal structure
20 comprising a three dimensional time series distribution composed
21 of a highly sparse data set. As used herein, the term "random"
22 (or "randomness") is defined in terms of a "random process" as
23 measured by the probability distribution model used, namely a
24 nearest-neighbor stochastic (Poisson) process. Thus, pure
25 randomness, pragmatically speaking, is herein considered to be a

1 time series distribution for which no function, mapping or
2 relation can be constituted that provides meaningful insight into
3 the underlying structure of the distribution, but which at the
4 same time is not chaos.

5 (2) Description of the Prior Art

6 Recent research has revealed a critical need for highly
7 sparse data set time distribution analysis methods and apparatus
8 separate and apart from those adapted for treating large sample
9 distributions. This is particularly the case in applications
10 such as naval sonar systems, which require that input time series
11 signal distributions be classified according to their structure,
12 i.e., periodic, transient, random or chaotic. It is well known
13 that large sample methods often fail when applied to small sample
14 distributions, but that the same is not necessarily true for
15 small sample methods applied to large data sets. Very small data
16 set distributions may be defined as those with less than about
17 ten (10) to twenty-five (25) measurement (data) points. Such
18 data sets can be analyzed mathematically with certain
19 nonparametric discrete probability distributions, as opposed to
20 large-sample methods, which normally employ continuous
21 probability distributions (such as the Gaussian).

22 The probability theory discussed herein and utilized by the
23 present invention is well known. It may be found, for example,
24 in works such as P.J. Hoel et al., Introduction to the Theory of

1 Probability, Houghton-Mifflin, Boston, MA, 1971, which is hereby
2 incorporated herein by reference.

3 Also, as will appear more fully below, it has been found to
4 be important to treat white noise signals themselves as the time
5 series signal distribution to be analyzed, and to identify the
6 characteristics of that distribution separately. This aids in
7 the detection and appropriate processing of received signals in
8 numerous data acquisition contexts, not the least of which
9 include naval sonar applications. Accordingly, it will be
10 understood that prior analysis methods and apparatus analyze
11 received time series data distributions from the point of view of
12 attempting to find patterns or some other type of correlated data
13 therein. Once such a pattern or correlation is located, the
14 remainder of the distribution is simply discarded as being noise.
15 It is believed that the present invention will be useful in
16 enhancing the sensitivity of present analysis methods, as well as
17 being useful on its own.

18 Various aspects related to the present invention are
19 discussed in the following exemplary patents:

20 U.S. Patent No. 6,068,659, issued May 30, 2000, to Francis
21 J. O'Brien, Jr., discloses a method for measuring and recording
22 the relative degree of pical density, congestion, or crowding of
23 objects dispersed in a three-dimensional space. A Population
24 Density Index is obtained for the actual conditions of the
25 objects within the space as determined from measurements taken of

1 the objects. The Population Density Index is compared with values
2 considered as minimum and maximum bounds, respectively, for the
3 Population Density Index values. The objects within the space are
4 then repositioned to optimize the Population Density Index, thus
5 optimizing the layout of objects within the space.

6 U.S. Patent No. 5,506,817, issued April 9, 1996, to Francis
7 J. O'Brien, Jr., discloses an adaptive statistical filter system
8 for receiving a data stream comprising a series of data values
9 from a sensor associated with successive points in time. Each
10 data value includes a data component representative of the motion
11 of a target and a noise component, with the noise components of
12 data values associated with proximate points in time being
13 correlated. The adaptive statistical filter system includes a
14 prewhitener, a plurality of statistical filters of different
15 orders, stochastic decorrelator and a selector. The prewhitener
16 generates a corrected data stream comprising corrected data
17 values, each including a data component and a time-correlated
18 noise component. The plural statistical filters receive the
19 corrected data stream and generate coefficient values to fit the
20 corrected data stream to a polynomial of corresponding order and
21 fit values representative of the degree of fit of corrected data
22 stream to the polynomial. The stochastic decorrelator uses a
23 spatial Poisson process statistical significance test to
24 determine whether the fit values are correlated. If the test
25 indicates the fit values are not randomly distributed, it

1 generates decorrelated fit values using an autoregressive moving
2 average methodology which assesses the noise components of the
3 statistical filter. The selector receives the decorrelated fit
4 values and coefficient values from the plural statistical filters
5 and selects coefficient values from one of the filters in
6 response to the decorrelated fit values. The coefficient values
7 are coupled to a target motion analysis module which determines
8 position and velocity of a target.

9 U.S. Patent No. 6,466,516 B1, issued October, 15, 2002, to
10 O'Brien, Jr. et al., discloses a method and apparatus for
11 automatically characterizing the spatial arrangement among the
12 data points of a three-dimensional time series distribution in a
13 data processing system wherein the classification of said time
14 series distribution is required. The method and apparatus
15 utilize grids in Cartesian coordinates to determine (1) the
16 number of cubes in the grids containing at least one input data
17 point of the time series distribution; (2) the expected number of
18 cubes which would contain at least one data point in a random
19 distribution in said grids; and (3) an upper and lower
20 probability of false alarm above and below said expected value
21 utilizing a discrete binomial probability relationship in order
22 to analyze the randomness characteristic of the input time series
23 distribution. A labeling device also is provided to label the
24 time series distribution as either random or nonrandom, and/or
25 random or nonrandom within what probability, prior to its output

1 from the invention to the remainder of the data processing system
2 for further analysis.

3 U.S. Patent No. 6,397,234 B1, issued May 28, 2002, to
4 O'Brien, Jr. et. al., discloses a method and apparatus for
5 automatically characterizing the spatial arrangement among the
6 data points of a time series distribution in a data processing
7 system wherein the classification of said time series
8 distribution is required. The method and apparatus utilize a
9 grid in Cartesian coordinates to determine (1) the number of
10 cells in the grid containing at least-one input data point of the
11 time series distribution; (2) the expected number of cells which
12 would contain at least one data point in a random distribution in
13 said grid; and (3) an upper and lower probability of false alarm
14 above and below said expected value utilizing a discrete binomial
15 probability relationship in order to analyze the randomness
16 characteristic of the input time series distribution. A labeling
17 device also is provided to label the time series distribution as
18 either random or nonrandom, and/or random or nonrandom.

19 U.S. Patent No. 6,597,634 B1, issued July 22, 2003, to
20 O'Brien, Jr. et al., discloses a signal processing system to
21 processes a digital signal converted from to an analog signal,
22 which includes a noise component and possibly also an information
23 component comprising small samples representing four mutually
24 orthogonal items of measurement information representable as a
25 sample point in a symbolic Cartesian four-dimensional spatial

1 reference system. An information processing sub-system receives
2 said digital signal and processes it to extract the information
3 component. A noise likelihood determination sub-system receives
4 the digital signal and generates a random noise assessment of
5 whether or not the digital signal comprises solely random noise,
6 and if not, generates an assessment of degree-of-randomness. The
7 information processing system is illustrated as combat control
8 equipment for undersea warfare, which utilizes a sonar signal
9 produced by a towed linear transducer array, and whose mode
10 operation employs four mutually orthogonal items of measurement
11 information.

12 The above prior art does not disclose a method which
13 utilizes more than one statistical test for characterizing the
14 spatial arrangement among the data points of a three dimensional
15 time series distribution of sparse data in order to maximize the
16 likelihood of a correct decision in processing batches of the
17 sparse data in real time operating submarine systems and/or other
18 contemplated uses.

19

20 SUMMARY OF THE INVENTION

21 Accordingly, it is an object of the invention to provide a
22 dual method comprising automated measurement of the three
23 dimensional spatial arrangement among a very small number of
24 points, objects, measurements or the like whereby an

1 ascertainment of the noise degree (i.e., randomness) of the time
2 series distribution may be made.

3 It also is an object of the invention to provide a dual
4 method and apparatus useful in naval sonar, radar and lidar and
5 in aircraft and missile tracking systems, which require acquired
6 signal distributions to be classified according to their
7 structure (i.e., periodic, transient, random, or chaotic) in the
8 processing and use of those acquired signal distributions as
9 indications of how and from where they were originally generated.

10 Further, it is an object of the invention to provide a dual
11 method and apparatus capable of labeling a three dimensional time
12 series distribution with (1) an indication as to whether or not
13 it is random in structure, and (2) an indication as to whether or
14 not it is random within a probability of false alarm of a
15 specific randomness calculation.

16 These and other objects, features, and advantages of the
17 present invention will become apparent from the drawings, the
18 descriptions given herein, and the appended claims. However, it
19 will be understood that above listed objects and advantages of
20 the invention are intended only as an aid in understanding
21 certain aspects of the invention, are not intended to limit the
22 invention in any way, and do not form a comprehensive or
23 exclusive list of objects, features, and advantages.

24 Accordingly, the present invention provides a two-stage
25 method for characterizing a spatial arrangement among data points

1 for each of a plurality of three-dimensional time series
2 distributions comprising a sparse number of the data points.
3 The method may comprise one or more steps such as, for instance,
4 creating a first virtual volume containing a first three-
5 dimensional time series distribution of the data points to be
6 characterized and then subdividing the first virtual volume into
7 a plurality k of three-dimensional volumes such that each of the
8 plurality k of three-dimensional volumes have the same shape and
9 size.

10 A first stage characterization of the spatial arrangement of
11 the first three-dimensional time series distribution of the data
12 points may comprise the steps of determining a statistically
13 expected proportion Θ of the plurality k of three-dimensional
14 volumes containing at least one of the data points for a random
15 distribution of the data points such that $k * \Theta$ is a
16 statistically expected number M of the plurality k of three-
17 dimensional volumes which contain at least one of the data points
18 if the first three-dimensional time series distribution is
19 characterized as random. Other steps may comprise counting a
20 number m of the plurality k of three-dimensional volumes which
21 actually contain at least one of the data points in the first
22 three-dimensional time series distribution in any particular
23 sample. The method comprises statistically determining an upper
24 random boundary greater than M and a lower random barrier less
25 than M such that if the number m is between the upper random

1 barrier and the lower random barrier then the first time series
2 distribution is characterized as random in structure during the
3 first stage characterization.

4 A second stage characterization of the first three-
5 dimensional time series distribution of the data points may
6 comprise the steps of determining when Θ is less than a pre-
7 selected value, and then utilizing a Poisson distribution to
8 determine a mean of the data points. If Θ is greater than the
9 pre-selected value, then the method may comprise utilizing a
10 binomial distribution to determine a mean of the data points.
11 Additional steps may comprise computing a probability p from the
12 mean so determined based on whether Θ is greater than or less
13 than the pre-selected value. Other steps may comprise
14 determining a false alarm probability α based on a total number
15 of the plurality k of three-dimensional volumes for the first
16 three-dimensional time series distribution of the data points to
17 be characterized. The method may comprise comparing p with α to
18 determine whether to characterize the sparse data as noise or
19 signal during the second stage characterization.

20 The first stage characterization of the first three-
21 dimensional time series distribution of the data points is
22 compared with the second stage characterization of the first
23 three-dimensional time series distribution of the data points to
24 improve the overall accuracy of the characterization.

1 If the first stage characterization of the first three-
2 dimensional time series distribution of the data points indicates
3 a random distribution and the second stage characterization of
4 the first three-dimensional time series distribution of the data
5 points indicates a signal, then the method may comprise
6 continuing to process the data points.

7 If the first stage characterization of the first three-
8 dimensional time series distribution of the data points indicates
9 a random distribution and the second stage characterization of
10 the first three-dimensional time series distribution of the data
11 points indicates a random distribution, then the first three-
12 dimensional time series distribution of the data points as random
13 with a higher confidence level than in a single stage
14 characterization.

15 The method may continue for characterizing each of the
16 plurality of three-dimensional time series distribution of data
17 points.

18 In a preferred embodiment, the random process (white noise)
19 detection subsystem includes an input for receiving a three-
20 dimensional time series distribution of data points expressed in
21 Cartesian coordinates. This set of data points will be
22 characterized by no more than a maximum number of points having
23 values (amplitudes) between maximum and minimum values received
24 within a preselected time interval. A hypothetical
25 representation of a white noise time series signal distribution

1 in Cartesian space is illustratively shown in FIG. 1. The
2 invention is specifically adapted to analyze both selected
3 portions of such time series distributions, and the entirety of
4 the distribution depending upon the sensitivity of the randomness
5 determination, which is required in any particular instance.

6 The input time series distribution of data points is
7 received by a display/operating system adapted to accommodate a
8 pre-selected number of data points N in a pre-selected time
9 interval Δt and dispersed in three-dimensional space along with
10 a first measure referred to as Y with magnitude $\Delta Y = \max(Y) - \min(Y)$,
11 and a second measure referred to as Z with magnitude
12 $\Delta Z = \max(Z) - \min(Z)$. The display/operating system then creates a
13 virtual volume around the input data distribution and divides the
14 virtual volume into a grid consisting of cubic cells each of
15 equal enclosed volume. Ideally, the cells fill the entire
16 virtual volume, but if they do not, the unfilled portion of the
17 virtual volume is disregarded in the randomness determination.

18 An analysis device then examines each cell to determine
19 whether or not one or more of the data points of the input time
20 series distribution are located therein. Thereafter, a counter
21 calculates the number of occupied cells. Also, the number of
22 cells which would be expected to be occupied in the grid for a
23 totally random distribution is predicted by a computer device
24 according to known Poisson probability process theory and
25 binomial Theorem equations. In addition, the statistical bounds

1 of the predicted value are calculated based upon known discrete
2 binomial criteria.

3 A comparator is then used to determine whether or not the
4 actual number of occupied cells in the input time series
5 distribution is the same as the predicted number of cells for a
6 random distribution. If it is, the input time series
7 distribution is characterized as random. If it is not, the input
8 time series distribution is characterized as nonrandom.

9 Thereafter, the characterized time series distribution is
10 labeled as random or nonrandom, and/or as random or nonrandom
11 within a pre-selected probability rate of the expected randomness
12 value prior to being output back to the remainder of the data
13 processing system. In the naval sonar signal processing context,
14 this output either alone, or in combination with overlapping
15 similarly characterized time series signal distributions, will be
16 used to determine whether or not a particular group of signals is
17 white noise. If that group of signals is white noise, it
18 commonly will be deleted from further data processing. Hence, it
19 is contemplated that the present invention, which is not
20 distribution dependent in its analysis as most prior art methods
21 of signal analysis are, will be useful as a filter or otherwise
22 in conjunction with current data processing methods and
23 equipment.

24 In the above regards, it should be understood that the
25 statistical bounds of the predicted number of occupied cells in a

1 random distribution (including cells occupied by mere chance)
2 mentioned above may be determined by a second calculator device
3 using a so-called probability of false alarm rate. In this case,
4 the actual number of occupied cells is compared with the number
5 of cells falling within the statistical boundaries of the
6 predicted number of occupied cells for a random distribution in
7 making the randomness determination. This alternative embodiment
8 of the invention has been found to increase the probability of
9 being correct in making a randomness determination for any
10 particular time series distribution of data points by as much as
11 60%.

12 The above and other novel features and advantages of the
13 invention, including various novel details of construction and
14 combination of parts will now be more particularly described with
15 reference to the accompanying drawings and pointed out by the
16 claims. It will be understood that the particular device and
17 method embodying the invention is shown and described herein by
18 way of illustration only, and not as limitations on the
19 invention. The principles and features of the invention may be
20 employed in numerous embodiments without departing from the scope
21 of the invention in its broadest aspects.

22

23 BRIEF DESCRIPTION OF THE DRAWINGS

24 Reference is made to the accompanying drawings in which is
25 shown an illustrative embodiment of the apparatus and method of

1 the invention, from which its novel features and advantages will
2 be apparent to those skilled in the art, and wherein:

3 FIG. 1 is a hypothetical depiction in Cartesian coordinates
4 of a representative white noise (random) time series signal
5 distribution;

6 FIG. 2 is a hypothetical illustrative representation of a
7 virtual volume in accordance with the invention divided into a
8 grid of cubic cells each having a side of length δ , and an area
9 of δ^3 ;

10 FIG. 3 is a block diagram representatively illustrating the
11 method steps of the invention;

12 FIG. 4 is a block diagram representatively illustrating an
13 apparatus in accordance with the invention; and

14 FIG. 5 is a table showing an illustrative set of discrete
15 binomial probabilities for the randomness of each possible number
16 of occupied cells of a particular time series distribution within
17 a specific probability of false alarm rate of the expected
18 randomness number.

20 DESCRIPTION OF THE PREFERRED EMBODIMENT

21 Referring now to the drawings, a preferred embodiment of the
22 dual method of the invention will be presented first from a
23 theoretical perspective, and thereafter, in terms of a specific
24 example. In this regard, it is to be understood that all data
25 points are herein assumed to be expressed and operated upon by

1 the various apparatus components in a Cartesian coordinate
2 system. Accordingly, all measurement, signal and other data
3 input existing in terms of other coordinate systems is assumed to
4 have been re-expressed in a Cartesian coordinate system prior to
5 its input into the inventive apparatus or the application of the
6 inventive method thereto.

7 The invention starts from the preset capability of a
8 display/operating system 8 (FIG. 4) to accommodate a set number
9 of data points N in a given time interval Δt . The data points
10 are dispersed in three-dimensional space with a first measure
11 referred to as Y with magnitude $\Delta Y = \max(Y) - \min(Y)$, and a second
12 measure referred to as Z with magnitude $\Delta Z = \max(Z) - \min(Z)$. A
13 representation of a three-dimensional time series distribution of
14 random data points 4 is shown in FIG. 1. A subset 4a of this
15 overall time series data distribution would normally be selected
16 for analysis of its signal component distribution by this
17 invention.

18 For purposes of mathematical analysis of the signal
19 components, it is assumed that the product/quantity given by
20 $\Delta t * \Delta Y * \Delta Z = [\max(t) - \min(t)] * [\max(Y) - \min(Y)] * [\max(Z) - \min(Z)]$ will define the
21 virtual volume 4b, illustrated as containing the subset 4a, with
22 respect to the quantities in the analysis subsystem. The sides
23 of virtual volume are drawn parallel to the time axis and other
24 axes as shown. Then, for substantially the total volume of the
25 display region, a Cartesian partition is superimposed on the

1 region with each partition being a small cube of sides δ (see,
 2 FIG. 2). The measure of δ will be defined herein as:

$$3 \quad \delta = \left(\frac{\Delta t * \Delta Y * \Delta Z}{k} \right)^{\frac{1}{3}} \quad (1)$$

4 The quantity k represents the total number of small cubes of
 5 volume δ^3 created in the volume $\Delta t * \Delta Y * \Delta Z$. Other than full cubes
 6 are ignored in the analysis. The quantity of such cubes with
 7 which it is desired populate the display region is determined
 8 using the following relationship, wherein N is the maximum number
 9 of data points in the time series distribution, Δt , ΔY and ΔZ
 10 are the Cartesian axis lengths, and the side lengths of each of
 11 the cubes is δ :

$$12 \quad k_i = \text{int} \left(\frac{\Delta t}{\delta_i} \right) * \text{int} \left(\frac{\Delta Y}{\delta_i} \right) * \text{int} \left(\frac{\Delta Z}{\delta_i} \right), \quad (2)$$

13 where int is the integer operator,

$$14 \quad \delta_i = \sqrt[3]{\frac{\Delta t * \Delta Y * \Delta Z}{k_0}}, \text{ and}$$

$$15 \quad k_0 = \begin{cases} k_1 & \text{if } |N - k_1| \leq |N - k_2| \\ k_2 & \text{otherwise} \end{cases}$$

16 where

$$17 \quad k_1 = \left[\text{int} \left(N^{\frac{1}{3}} \right) \right]^3$$

$$18 \quad k_2 = \left[\text{int} \left(N^{\frac{1}{3}} \right) + 1 \right]^3 ;$$

$$k_{II} = \text{int}\left(\frac{\Delta t}{\delta_{II}}\right) * \text{int}\left(\frac{\Delta Y}{\delta_{II}}\right) * \text{int}\left(\frac{\Delta Z}{\delta_{II}}\right) \quad (3)$$

where

$$\delta_{II} = \sqrt[3]{\frac{\Delta t * \Delta Y * \Delta Z}{N}},$$

$$\therefore k = \begin{cases} k_I & \text{if } K_I > K_{II} \\ k_{II} & \text{if } K_I < K_{II} \\ \max(k_I, k_{II}) & \text{if } K_I = K_{II} \end{cases} \quad (4)$$

where

$$K_I = \frac{k_I}{\Delta t * \Delta Y * \Delta Z} \delta_I^3 \leq 1 \quad \text{and}$$

$$K_{II} = \frac{k_{II}}{\Delta t * \Delta Y * \Delta Z} \delta_{II}^3 \leq 1.$$

It is to be noted that in cases with very small amplitudes, it may occur that $\text{int}(\Delta Y/\delta_I) \leq 1$, $\text{int}(\Delta Y/\delta_{II}) \leq 1$, $\text{int}(\Delta Z/\delta_I) \leq 1$, or $\text{int}(\Delta Z/\delta_{II}) \leq 1$. In such cases, the solution is to round off either quantity to the next highest value (i.e., ≥ 2). This weakens the theoretical approach, but it allows for practical measurements to be made.

As an example of determining k , assume Δt (or N) = 30, ΔY = 20 and ΔZ = 9, then k = 30 (from equations (2) through (4)) and δ = 5.65 (from equation (1)). In essence, therefore, the above relation defining the value k selects the number of cubes having sides of length δ and volume δ^3 , which fill up the total space $\Delta t * \Delta Y * \Delta Z$ to the greatest extent possible, i.e., $k * \delta^3 \approx \Delta t * \Delta Y * \Delta Z$.

1 From the selected partitioning parameter k , the region
2 (volume) $\Delta t * \Delta Y * \Delta Z$ is carved up into k cubes, with the sides of
3 each cube being δ as defined above. In other words, the
4 horizontal (or time) axis is marked off into intervals, exactly
5 $\text{int}(\Delta t / \delta)$ of them, so that the time axis has the following
6 arithmetic sequence of cuts (assuming that the time clock starts
7 at $\Delta t = 0$):

$$8 \qquad 0, \delta, 2\delta, \dots, \text{int}(\Delta t / \delta) * \delta$$

9 Likewise, the vertical (or first measurement) axis is cut up
10 into intervals, exactly $\text{int}(\Delta Y / \delta)$ of them, so that the vertical
11 axis has the following arithmetic sequence of cuts:

$$12 \qquad \min(Y), \min(Y) + \delta, \dots, \min(Y) + \text{int}(\Delta Y / \delta) * \delta = \max(Y),$$

13 where \min is the minimum operator and \max is the maximum
14 operator.

15 Similarly, the horizontal plane (or second measurement) axis
16 is cut up into intervals, exactly $\text{int}(\Delta Z / \delta)$ of them, so that this
17 horizontal plane axis has the following arithmetic sequence of
18 cuts:

$$19 \qquad \min(Z), \min(Z) + \delta, \dots, \min(Z) + \text{int}(\Delta Z / \delta) * \delta = \max(Z).$$

20 Based on the Poisson point process theory for a measurement
21 set of data in a time interval Δt of measurements of magnitudes
22 ΔY and ΔZ , that data set is considered to be purely random (or
23 "white noise") if the number of partitions k are nonempty (i.e.,

1 contain at least one data point of the time series distribution
2 thereof under analysis) to a specified degree. The expected
3 number of nonempty partitions in a random distribution is given
4 by the relationship:

$$k^*\Theta = k^*(1 - e^{-N/k}) \quad (5)$$

6 where the quantity Θ is the expected proportion of nonempty
7 partitions in a random distribution and N/k is "the parameter of
8 the spatial Poisson process" corresponding to the average number
9 of points observed across all three-dimensional subspace
10 partitions.

11 The boundary, above and below $k^*\Theta$, attributable to random
12 variation and controlled by a false alarm rate is the so-called
13 "critical region" of the test. The quantity Θ not only
14 represents (a) the expected proportion of nonempty cubic
15 partitions in a random distribution, but also (b) the probability
16 that one or more of the k cubic partitions is occupied by pure
17 chance, as is well known to those in the art. The boundaries of
18 the parameter $k^*\Theta$ comprising random process are determined in
19 the following way.

20 Let M be a random variable representing the integer number
21 of occupied cubic partitions as illustratively shown in FIG. 2.
22 Let m be an integer (sample) representation of M . Let m_1 be the
23 quantity forming the lower random boundary of the statistic $k^*\Theta$
24 given by the binomial criterion:

$$P(M \leq m) \leq \frac{\alpha_0}{2}, \min\left(\frac{\alpha}{2} - \frac{\alpha_0}{2}\right)$$

where,

$$P(M \leq m) = \sum_{m=0}^{m_1} B(m; k, \Theta) . \quad (6)$$

$B(m; k, \Theta)$ is the binomial probability function given as:

$$B(m; k, \Theta) = \binom{k}{m} (\Theta)^m (1 - \Theta)^{k-m}$$

where $\binom{k}{m}$ is the binomial coefficient,

$$\binom{k}{m} = \frac{k!}{m!(k-m)!}, \text{ and} \quad (6A)$$

$$\sum_{m=0}^{m=k} B(m; k, \Theta) = 1.0 .$$

The quantity α_0 is the probability of coming closest to an exact value of the pre-specified false alarm probability α , and m_1 is the largest value of m such that $P(M \leq m) \leq \alpha_0/2$. It is an objective of this method to minimize the difference between α and α_0 . The recommended probability of false alarm (PFA) values for differing values of spatial subsets k , and based on commonly accepted levels of statistical precision, are as follows:

PFA(α)	k
0.01	$k \geq 25$
0.05	$k < 25$

The upper boundary of the random process is called m_2 , and is determined in a manner similar to the determination of m_1 .

Thus, let m_2 be the upper random boundary of the statistic $k^* \oplus$ given by:

$$P(M \geq m) \leq \frac{\alpha_0}{2}, \min\left(\frac{\alpha}{2} - \frac{\alpha_0}{2}\right)$$

where

$$P(M \geq m) = \sum_{m=m_1}^k B(m; k, \Theta) \leq \alpha_o / 2$$

or

$$P(M \geq m) = 1 - \sum_{m=0}^{m_2-1} B(m; k, \Theta) \leq \alpha_o / 2. \quad (7)$$

The value α_0 is the probability coming closest to an exact value of the pre-specified false alarm probability α , and m_2 is the largest value of m such that $P(M \geq m) \leq \alpha_0/2$. It is an objective of the invention to minimize the difference between α and α_0 .

Hence, the subsystem determines if the signal structure contains m points within the "critical region" warranting a determination of "non-random", or else "random" is the determination, with associated PFA of being wrong in the decision when "random" is the decision.

The subsystem also assesses the random process hypothesis by testing:

1
$$H_0: \hat{P} = \Theta(\text{NOISE})$$

2
$$H_1: \hat{P} \neq \Theta(\text{SIGNAL} + \text{NOISE}),$$

3 where $\hat{P} = m/k$ is the sample proportion of signal points contained
 4 in the k sub-region partitions of the space $\Delta t * \Delta Y * \Delta Z$ observed in
 5 a given time series. As noted above, FIG. 1 shows what a
 6 hypothetical white noise (random) distribution looks like in
 7 Cartesian time-space.

8 Thus, if $\Theta \approx \hat{P} = m/k$, the observed distribution conforms to a
 9 random distribution corresponding to "white noise".

10 The estimate for the proportion of k cells occupied by N
 11 measurements (\hat{P}) is developed in the following manner. Let each
 12 of the k cubes with sides of length δ be denoted by C_{hij} , and the
 13 number of objects observed in each C_{hij} cube be denoted $\text{card}(C_{hij})$
 14 where card means "cardinality" or subset count. C_{hij} is labeled
 15 in an appropriate manner to identify each and every cube in the
 16 three space. Using the example given previously with $N = \Delta t = 30$,
 17 $\Delta Y = 20$, $\Delta Z = 9$ and $k = 30 = 5 * 3 * 2$, the cubes may be labeled using the
 18 index h running from 1 to 5, the index i running from 1 to 3 and
 19 the index j running from 1 to 2 (see FIG. 2).

20 Next, to continue the example for $k = 30$ shown in FIG. 2,
 21 define the following cube counting scoring scheme for the $5 * 3 * 2$
 22 partitioning comprising whole cube subsets:

$$X_{hij} = \begin{cases} 1 & \text{if } \text{card}(C_{hij}) > 0; h=1 \text{ to } 5, i=1 \text{ to } 3, j=1 \text{ to } 2 \\ 0 & \text{if } \text{card}(C_{hij}) = 0; h=1 \text{ to } 5, i=1 \text{ to } 3, j=1 \text{ to } 2 \end{cases}$$

Thus, X_{hij} is a dichotomous variable taking on the individual values of 1 if a cube C_{hij} has one or more objects present, and a value of 0 if the cube is empty.

Then calculate the proportion of 30 cells occupied in the partition region:

$$\hat{P} = \frac{1}{30} \sum_{j=1}^2 \sum_{i=1}^3 \sum_{h=1}^5 X_{hij} .$$

The generalization of this example to any sized table is obvious and within the scope of the present invention. For the general case, it will be appreciated that, for the statistics X_{hij} and C_{hij} , the index h runs from 1 to $\text{int}(\Delta t/\delta)$, the index i runs from 1 to $\text{int}(\Delta Y/\delta)$ and the index j runs from 1 to $\text{int}(\Delta Z/\delta)$.

In addition, a conjoint, confirmatory measure useful in the interpretation of outcomes is the R ratio, defined as the ratio of observed to expected occupancy rates:

$$R = \frac{m}{k * \Theta} = \frac{\hat{P}}{\Theta} \quad (8)$$

The range of values for R indicate:

$R < 1$, clustered distribution

$R = 1$, random distribution; and

$R > 1$, uniform distribution.

The R statistic is used in conjunction with the formulation just described involving the binomial probability distribution

1 and false alarm rate in deciding to accept or reject the "white
2 noise" hypothesis. Its use is particularly warranted in very
3 small samples ($N < 25$). In actuality, R may never have a precise
4 value of 1. Therefore, a new novel method is employed for
5 determining randomness based on the R statistic of equation (8).

6 A rigorous statistical procedure has been developed to
7 determine whether the observed R -value is indicative of "noise"
8 or "signal". The procedure renders quantitatively the
9 interpretations of the R -value whereas the prior art has relied
10 primarily on intuitive interpretation or ad hoc methods, which
11 can be erroneous.

12 In this formulation, one of two statistical assessment tests
13 is utilized depending on the value of the parameter Θ .

14 If $\Theta \leq 0.10$, then a Poisson distribution is employed. To
15 apply the Poisson test, the distribution of the N sample points
16 is observed in the partitioned space. It will be appreciated
17 that a data sweep across all cells within the space will detect
18 some of the squares being empty, some containing $k = 1$ points, k
19 $= 2$ points, $k = 3$ points, and so on. The number of points in
20 each k category is tabulated in a table such as follows:

Frequency Table of Cell Counts

k (number of cells with points)	N_k (number of points in k cells)
0	N_0
1	N_1
2	N_2
3	N_3
\vdots	\vdots
K	N_K

From this frequency table, two statistics are of interests for the Central Limit Theorem approximation:

The "total", $Y = \sum_{k=0}^K kN_k$, and (9)

the sample mean, $\mu_0 = \frac{\sum_{k=0}^K kN_k}{\sum_{k=0}^K N_k}$.

Then, if $\Theta \leq 0.10$, the following binary hypothesis is of interest:

$$\begin{aligned} H_0: \mu &= \mu_0(\text{NOISE}) \\ H_1: \mu &\neq \mu_0(\text{SIGNAL}) \end{aligned} \quad (10)$$

The Poisson test statistic, derived from the Central Limit Theorem, Eq. (9) is as follows:

$$Z_p = \frac{Y - N\mu_0}{\sqrt{N\mu_0}}, \quad (k > 25) \quad (11)$$

where

$$Y = \sum_{k=0}^K kN_k,$$

1 and N is the sample size. Then

2
$$\mu_0 = \frac{\sum_{k=0}^K kN_k}{\sum_{k=0}^K N_k}$$
 is the sample mean and sample variance. (It is

3 well known that $\mu = \sigma^2$ in a Poisson distribution).

4 The operator compares the value of Z_p against a probability
5 of False Alarm α . α is the probability that the null
6 hypothesis (NOISE) is rejected when the alternative (SIGNAL) is
7 the truth.

8 The probability of the observed value Z_p is calculated as:

9
$$p = P(|Z_p| \leq Z) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|Z_p|}^{|Z_p|} \exp(-.5x^2) dx \quad (12)$$

10 where $|x|$ means "absolute value" as commonly used in mathematics.

11 The calculation of Eq. 12 as known to those skilled in the
12 art, is performed in a standard finite series expansion.

13 On the other hand, if $\Theta > .10$, the invention dictates that
14 the following binary hypothesis set prevail:

15
$$\begin{aligned} H_0 : \mu &= k\theta(\text{NOISE}) \\ H_1 : \mu &= k\theta(\text{SIGNAL}) \end{aligned}$$

16 The following binomial test statistic is employed to test the
17 hypothesis:

18
$$Z_B = \frac{m \pm c - k\theta}{\sqrt{k\theta(1-\theta)}}$$

1 where $c = 0.5$ if $X < \mu$ and $c = -0.5$ if $X > \mu$ (Yates Continuity
2 correction factor used for discrete variables)

3 The quantities of Z_B have been defined previously.

4 The probability of the observed value Z_B is calculated as

5
$$p = P(|Z_B| \leq Z) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|Z_B|}^{|Z_B|} \exp(-.5x^2) dx$$

6 in a standard series expansion.

7 For either test statistic, Z_p or Z_B , the following decision
8 rule is used to compare the false alarm rate α with the observed
9 probability of the statistic, p :

10
$$\begin{aligned} &\text{if } p \geq \alpha \Rightarrow \text{NOISE} \\ &\text{If } p < \alpha \Rightarrow \text{SIGNAL} \end{aligned}$$

11 Thus, if the calculated probability value $p > \alpha$, then the
12 three-dimensional spatial distribution is deemed "noise";
13 otherwise the X-Y-Z data is characterized as "signal" by the
14 Rtest.

15

16 EXAMPLE

17 Having thus explained the theory of the invention, an
18 example thereof will now be presented for purposes of further
19 illustration and understanding (see, FIGS. 3 and 4). A value
20 for N is first selected, here $N = 30$ (step 100, FIG. 3). A time
21 series distribution of data points is then read into a
22 display/operating subsystem 8 adapted to accommodate a data set
23 of size N from data processing system 10 (step 102). Thereafter,

1 the absolute value of the difference between the largest and the
2 smallest data points for each measure, ΔY and, is determined by
3 a first comparator device 12 (step 104). In this example, it
4 will be assumed that $N = \Delta t = 30$ measurements with a measured
5 amplitudes of $\Delta Y = 20$ units and $\Delta Z = 9$ units. The N , ΔY and ΔZ
6 values are then used by window creating device 14 to create a
7 virtual volume in the display/operating system enclosing the
8 input time series distribution, the size of the volume so created
9 being $\Delta t * \Delta Y * \Delta Z = 5400$ units (step 106).

10 Thereafter, as described above, the virtual volume is
11 divided by the cube creating device 14 into a plurality k of
12 cubes C_{hij} (see FIG. 4), each cube having the same geometric shape
13 and enclosing an equal volume so as to substantially fill the
14 virtual volume containing the input time series distribution set
15 of data points (step 108). The value of k is established by the
16 relation given in equations (2) through (4):

$$k = \text{int}\left(\frac{\Delta t}{\delta}\right) * \text{int}\left(\frac{\Delta Y}{\delta}\right) * \text{int}\left(\frac{\Delta Z}{\delta}\right) = 5 * 3 * 2 = 30$$

$$\delta = \sqrt[3]{\frac{\Delta t * \Delta Y * \Delta Z}{k}} = 5.65.$$

19 Thus, the 5400 unit^3 space of the virtual volume is
20 partitioned into 30 cubes of side 5.65 so that the whole space is
21 filled ($k * \delta^3 = 5400$). The time-axis arithmetic sequence of
22 cuts are: 0, 5.65, ..., $\text{int}(\Delta t / \delta) \quad \delta = 5.65$. The Y amplitude

1 axis cuts are: $\min(Y)$, $\min(Y) + \delta$, ..., $\min(Y) + \text{int}(\Delta Y/\delta) * \delta =$
2 $\max(Y)$ and the Z amplitude axis cuts are: $\min(Z)$, $\min(Z) + \delta$,
3 ..., $\min(Z) + \text{int}(\Delta Z/\delta) * \delta = \max(Z)$.

4 Next, the probability false alarm rate is set at step 110
5 according to the value of k as discussed above. More
6 particularly, in this case $\alpha = 0.01$, and the probability of a
7 false alarm within the critical region is $\alpha/2 = 0.005$.

8 The randomness count is then calculated by first computing
9 device 16 at step 112 according to the relation of equation (5):

$$10 \quad k * \Theta = k * (1 - e^{-N/k}) = 30 * 0.632 \cong 18.96.$$

11 Therefore, the number of cubes expected to be non-empty in this
12 example, if the input time series distribution is random, is
13 about 19.

14 The binomial distribution discussed above is then calculated
15 by a second computing device 18 according to the relationships
16 discussed above (step 114, FIG. 3). Representative values for
17 this distribution are shown in FIG. 5 for each number of possible
18 occupied cells m for $k = 30$ and $\Theta = 0.632$.

19 The upper and lower randomness boundaries then are
20 determined, also by second calculating device 18. Specifically,
21 the lower boundary is calculated from FIG. 5 (step 116) from the
22 criterion $P(M \leq m) \leq \alpha_0/2$. Then, computing the binomial
23 probabilities results in $P(M \leq 11) = .00265$. Thus, the lower
24 bound is $m_1 = 11$.

1 The upper boundary, on the other hand, is the randomness
 2 boundary m_2 from the criterion $P(M \geq m) \leq \alpha_0/2$. Computing the
 3 binomial probabilities gives $P(M \geq 27) = .00435$; hence $m_2 = 27$ is
 4 taken as the upper bound (step 118). The probabilities necessary
 5 for this calculation also are shown in FIG. 5.

6 Therefore, the critical region is defined in this example as
 7 $m_1 \leq 11$, and $m_2 \geq 27$ (step 120).

8 The actual number of cells containing one or more data
 9 points of the time series distribution determined by
 10 analysis/counter device 20 (step 122, FIG. 3) is then used by
 11 divider 22 and a second comparator 24 in the determination of the
 12 randomness of the distribution (step 124, FIG. 3). Specifically,
 13 using $m = 18$ as an example, it will be seen that the sample
 14 statistic $\hat{P} = m/k = 0.600$, and that $R = \hat{P}/\Theta = 0.600/0.632 = 0.94$.

15 Branching to step 123 (FIG. 3) which the sparse data
 16 decision logic module performs, the R statistic value of 0.94 is
 17 evaluated statistically. A more precise indicator is obtained by
 18 applying the significance test in accord with the present
 19 invention, as described earlier. For this calculation, we note
 20 that $\theta = .632$, which invokes the Binomial probability model to
 21 test the hypothesis:

$$H_0: \mu = k\theta(\text{NOISE})$$

$$H_1: \mu = k\theta(\text{SIGNAL})$$

23 In this case, $k\theta = 18.96$. Thus, applying the Binomial test
 24 gives:

$$Z_B = \frac{m \pm c - k\theta}{\sqrt{k\theta(1-\theta)}}$$

$$= \frac{18 - .5 - 18.96}{\sqrt{30(.632)(1-.632)}} \approx -.55$$

The p value is computed to be:

$$p = P(|Z_B| \leq Z) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-|Z|}^{|Z|} \exp(-.5x^2) dx = .58$$

Since $p = .58$ and $\alpha = 0.1$, and since $p \geq \alpha$, we conclude that the R test shows the volumetric data to be random (NOISE only, with 99% certainty) with the value of $R = .93$ computed for this spatial distribution in 3D-space.

It is also worth noting in this regard that the total probability is $0.00265 + .00435 = .00700$, which is the probability of being wrong in deciding "random". This value is less than the probability of a false alarm, $PFA = 0.01$. Thus, the actual protection against an incorrect decision is much higher (by about 30%) than the *a priori* sampling plan specified.

Since $m = 18$ falls inside of the critical region, i.e., $m_1 \leq 18 \leq m_2$, the decision is that the data represent an essentially white noise distribution (step 126). Accordingly, since both methods yield consistent results the distribution is labeled at step 128 by the labeling device 26 as a noise distribution, and transferred back to the data processing system 10 for further processing. In the naval sonar situation having a spatial component, a signal distribution labeled as white noise would be

1 discarded by the processing system, but in some situations a
2 further analysis of the white noise nature of the distribution
3 would be possible. Similarly, the invention is contemplated to
4 be useful as an improvement on systems that look for patterns and
5 correlations among data points. For example, overlapping time
6 series distributions might be analyzed in order to determine
7 where a meaningful signal begins and ends.

8 It will be understood that many additional changes in the
9 details, materials, steps and arrangement of parts, which have
10 been herein described and illustrated in order to explain the
11 nature of the invention, may be made by those skilled in the art
12 within the principles and scope of the invention as expressed in
13 the appended claims.